Appendix A: Methodology

Appendix A: Methodology

I. The Gini Coefficient

The Gini coefficient can be defined in various ways but it is most easily explained by the Lorenz curve. The Lorenz curve is depicted by plotting the points of the household share of those households with incomes less than a certain level and their income share.[1] For example, in Figure A–1, P indicates that the poorest x per cent of households receive y per cent of the total income of all households.

By definition, the Lorenz curve must be below the diagonal (OB). When each household obtains the same level of income, the Lorenz curve coincides with the diagonal. Therefore, this diagonal is called the "egalitarian line" or "line of perfect equality." The Gini coefficient is defined as the ratio of the area between the Lorenz curve and the egalitarian line (area G in Figure A–1) to the triangle OAB in the same figure. The smaller the Gini coefficient is, income distribution becomes equal to a greater extent. If the Lorenz curve coincides with the egalitarian line, the Gini coefficient is zero. On the other hand, if only one household receives all income and

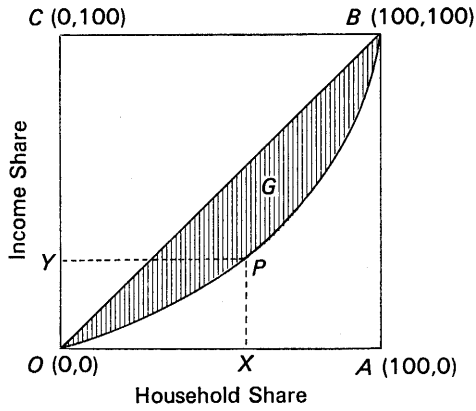

Figure A–1
Lorenz Curve

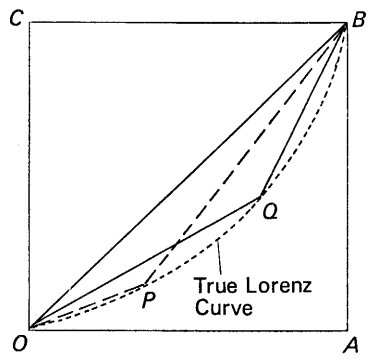

Figure A–2
Lorenz Curve and Income Brackets

all other households have no income, which is the case of the highest possible inequality, the Lorenz curve comes closest to the line *OAB* and the Gini coefficient is one, the highest. Thus as income inequality increases, the Gini coefficient increases from zero to one.

A. Direct Method

Income distribution data is often given by arbitrary income brackets. The problems in using this kind of data can be illustrated by the following example. For simplicity we assume that the income distribution data is given by two income classes: higher or lower than a certain income level Y. Now two income levels, Y_1 and Y_2 ($Y_1 < Y_2$), are arbitrarily chosen for Y. In Figure A–2 the dotted curve shows the true Lorenz curve and P and Q correspond to the income range, Y_1 and Y_2. The Lorenz curve which is observed with this income distribution data will look like the linked lines *OPB* and *OQB* in Figure A–2. By the direct method the Gini coefficient is calculated as the ratio of the area of the triangle *OPB* or *OQB* to the area of the triangle *OAB*. Thus the Gini coefficients resulting from the direct method would be different to each other even though the true Lorenz curve is the same.

B. Decile Method

One of the ways to avoid such a problem of the direct method is to fix the interval of the household group. One of these methods is the distribution of income by household decile. The household decile is given by dividing all households into ten groups with an equal number of households according to their income level. In this study we call the poorest 10 per cent the

bottom decile and the richest 10 per cent the top decile; the second decile, third decile, and so on in between are in ascending order from the bottom to the top according to the income level. To estimate the mean income of each decile we used the log-normal distribution and Pareto distribution.

The Log-normal Distribution

The log-normal distribution generally fits well to the lower and middle income classes. The log-normal distribution is expressed as:

$$[1] \quad P(y) = \int_0^y \frac{1}{\sqrt{2\pi}\sigma t} \exp\left\{\frac{(\ln t - m)^2}{2\sigma^2}\right\} dt,$$

where $P(y)$ is the proportion of households with income less than y, and m and σ are parameters.

An example of this curve is shown in Figure 3–4.

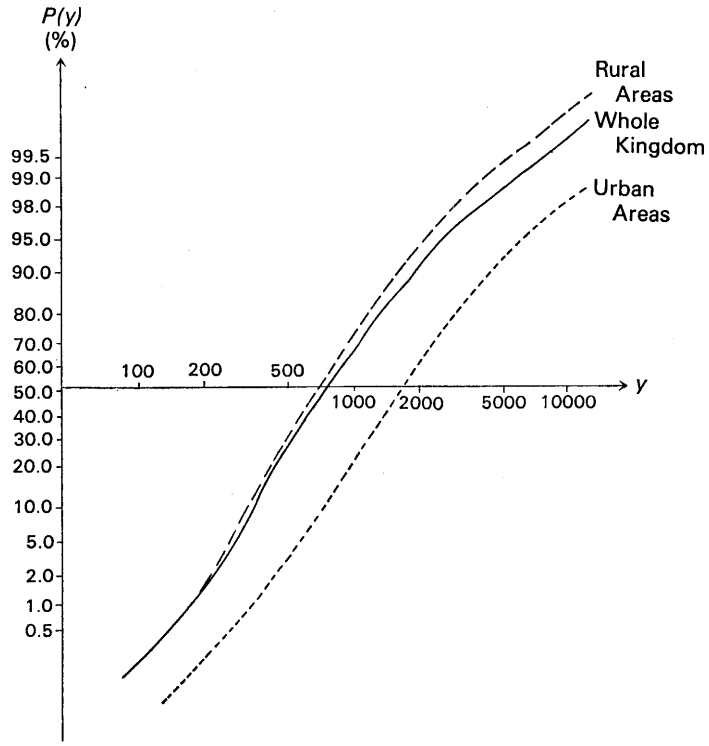

Figure A-3
Probit Graph, 1969
Source: Estimated from data in Meesook [36].

Equation [1] can be expressed simply as

$$[2] \quad P(y) = N\left(\frac{\ln y - m}{\sigma}\right),$$

where N is the cumulative form of the standard normal distribution.

Taking the inverse of equation [2],

$$[3] \quad N^{-1}(P(y)) = \frac{\ln y - m}{\sigma}.$$

This means that $N^{-1}(P(y))$ and $\ln y$ will be plotted on a straight line if a distribution follows the log-normal distribution. This graph is called the probit graph. To examine the log-normality the income distribution data for 1969, 1975, 1981, and 1986 are applied to this equation (Figures A–3 to A–6). As is often the case in other countries, the curves move upward at the lower income class and downward at the higher income class. It could be said, however, that except for these two classes the distribution follows

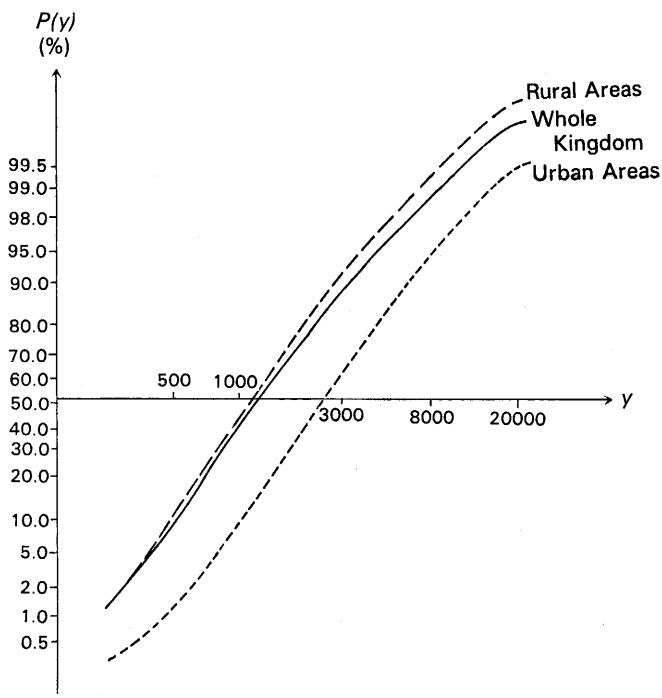

Figure A–4
Probit Graph, 1975
Source: Estimated from data tape of SES 1975/76.

the log-normal distribution. Next step is to estimate the parameters m and σ. By rearranging the equation [3] we have

$$[4] \quad \ln y = \sigma N^{-1}(P(y)) + m.$$

Then the parameters m and σ are estimated simply by applying *OLS* to this equation, excluding the data of the higher income classes.[2]

The results are shown in Table A–1. These results show that the log-normal distribution fits very well to all the cases. With these estimated parameters the income brackets for each decile group are calculated as:

$$Y_i = \exp\{\sigma N^{-1}(P_i) + m\}, \qquad P_i = 0.1,\ 0.2,\ 0.3,\ \ldots,\ 0.9.$$

Pareto Distribution

For the higher income classes the Pareto distribution generally proves to be very useful. Therefore we apply it to the higher income classes. The Pareto distribution is expressed as:

$$Q(y) = Ay^{-\alpha} \qquad A>0,\ \alpha>1,$$

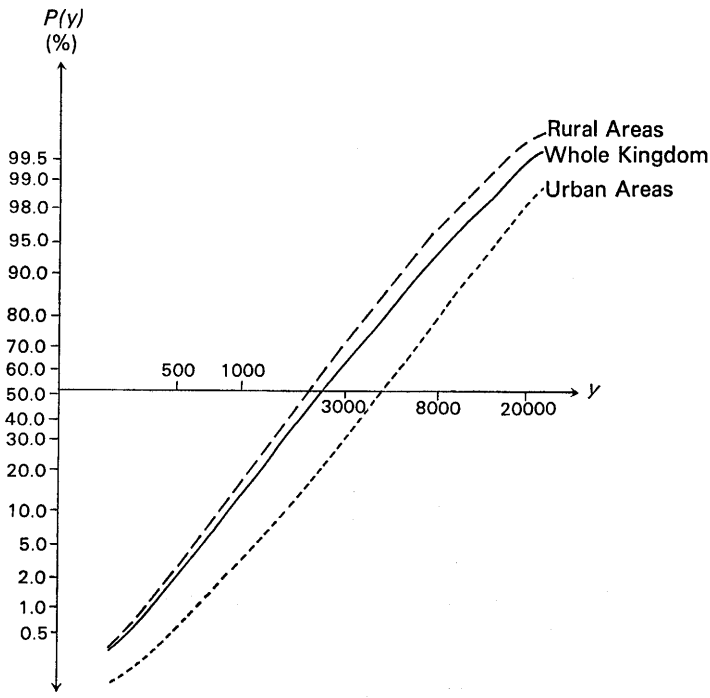

Figure A–5
Probit Graph, 1981
Source: Estimated from data tape of SES 1981.

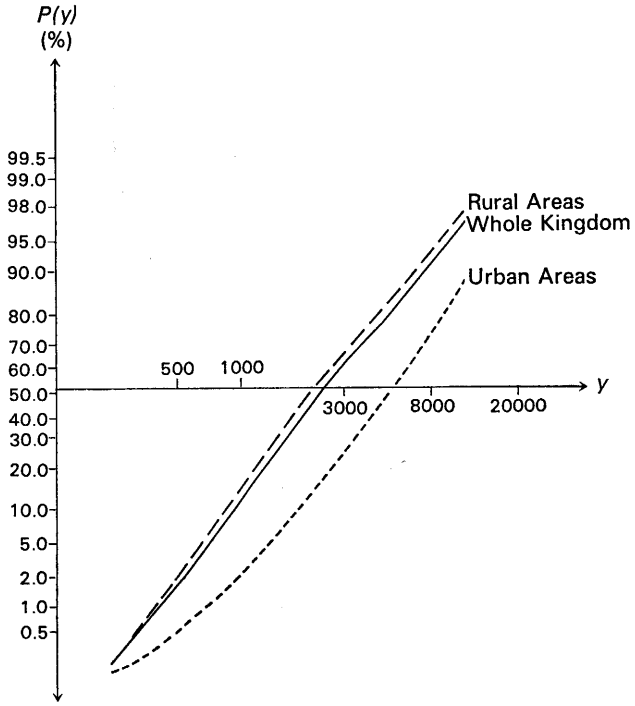

Figure A-6
Probit Graph, 1986
Source: Estimated from NSO [65].

Table A-1
Estimates of Log-normal Distribution

Year	Areas	Parameter Estimate		Adjusted R-Squared
		m	σ	
1969	Overall	9.1688 (730.4)	0.6689 (71.5)	0.9953
	Rural	9.0690 (715.2)	0.6278 (68.2)	0.9955
	Urban	9.9679 (401.2)	0.7382 (49.6)	0.9880
1975	Overall	7.2223 (661.9)	0.7248 (68.4)	0.9981
	Rural	7.1239 (630.8)	0.6892 (64.7)	0.9979
	Urban	7.8945 (253.2)	0.7393 (32.5)	0.9888
1981	Overall	7.7915 (917.2)	0.7722 (114.3)	0.9991
	Rural	7.6607 (727.9)	0.7284 (86.9)	0.9985
	Urban	8.3805 (341.0)	0.8104 (48.4)	0.9928
1986	Overall	7.8193 (572.1)	0.7668 (70.5)	0.9972
	Rural	7.7281 (639.0)	0.7302 (77.3)	0.9978
	Urban	8.4968 (240.1)	0.8316 (33.5)	0.9842

Source: Estimated by the author.
Note: Figures in the parentheses indicate t-value.

where $Q(y)=1-P(y)$.

By taking logarithm of this equation,

$$\ln Q(y) = \ln A - \alpha \cdot \ln y .$$

This means that if a distribution follows the Pareto distribution, $\ln Q(y)$ and $\ln y$ will be plotted on a straight line. This curve is the Pareto curve. This methodology is applied to the data of Thailand (Figures A–7 to A–10). From these figures it can be said that the Pareto distribution is a good approximation except for the lower income classes.

The parameters A and α are estimated by *OLS*. The results are shown

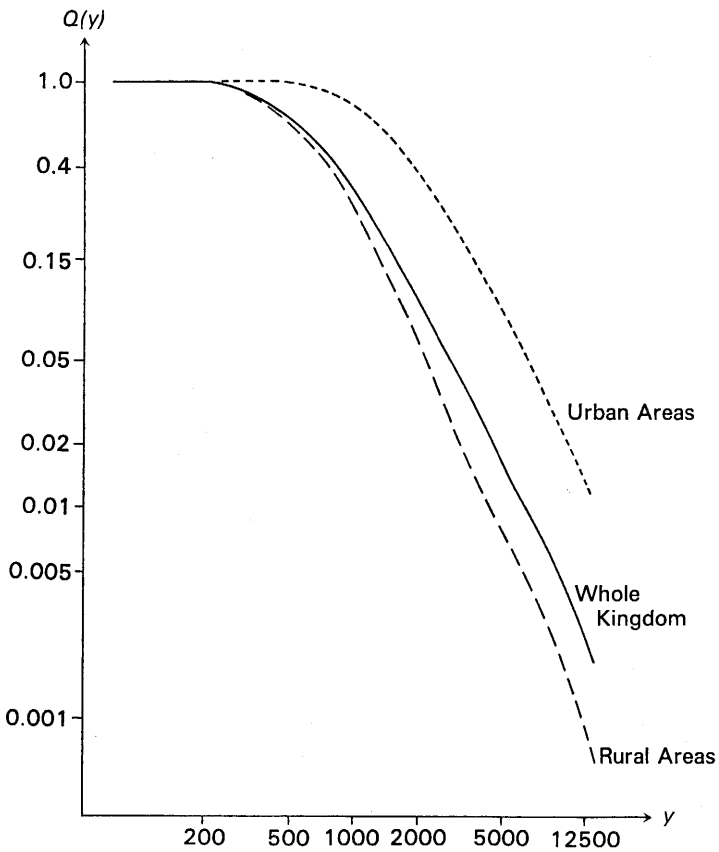

Figure A–7
Pareto Curve, 1969
Source: Estimated from data in Meesook [36].

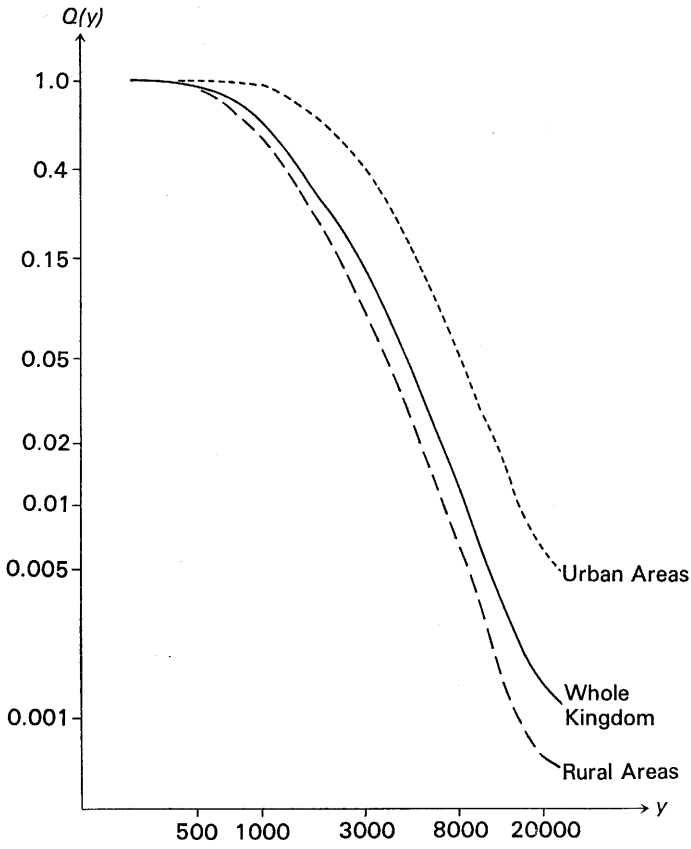

Figure A-8
Pareto Curve, 1975
Source: Estimated from data tape of SES 1975/76.

in Table A-2. With these estimated coefficients the income brackets of the top and ninth deciles are calculated as:

$$Y_i = \left(\frac{1-P_i}{A}\right)^{-1/\alpha} \qquad P_i = 0.8,\ 0.9.$$

By integration the mean income of these deciles is calculated as:

$$M_i = \begin{cases} \dfrac{10\alpha A^{1/\alpha}}{\alpha-1}\{(0.2)^{\alpha-1/\alpha}-(0.1)^{\alpha-1/\alpha}\} & \text{for the ninth decile.} \\ \dfrac{\alpha}{\alpha-1}(10A)^{1/\alpha} & \text{for the top decile.} \end{cases}$$

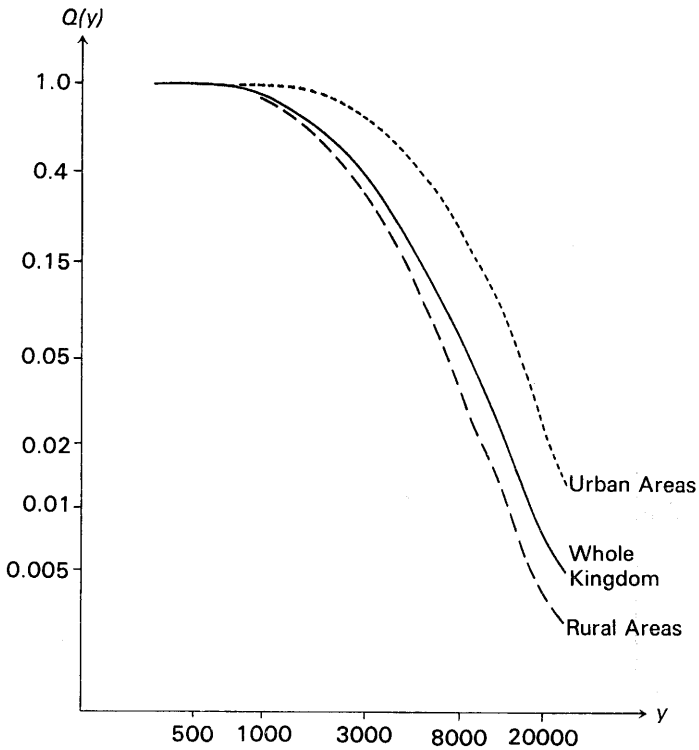

Figure A–9
Pareto Curve, 1981
Source: Estimated from data tape of SES 1981.

We use this mean income for the top decile. If the Pareto distribution fits well to the ninth decile we adopt the mean income of the ninth decile of this formula. If not, we choose the income range estimated by the log-normal distribution or the Pareto distribution, whichever proves to be better. For the lower deciles we use the arithmetic mean of the income interval as the mean income of the decile. Instead of the arithmetic mean, some scholars use the geometric mean. The estimates of this method are usually lower than our method. Therefore, the income inequality estimated by this method tends to be slightly bigger than our method.

Formula of the Decile Method

The decile method of estimating the Gini coefficient used in this study is expressed simply as follows:

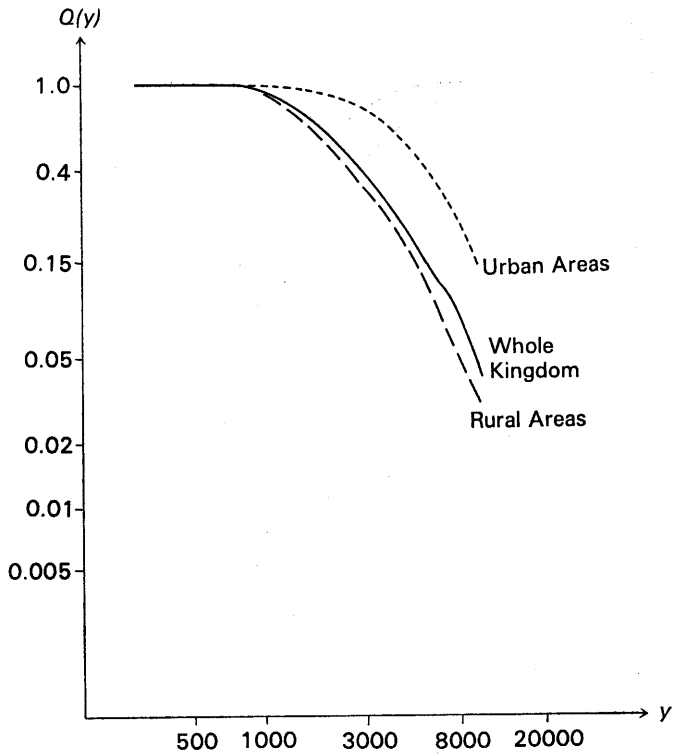

Figure A–10
Pareto Curve, 1986
Source: Estimated from NSO [65].

$$\text{Gini} = \left\{\sum_{i=1}^{10}\left(i-\frac{11}{2}m_i\right)\right\}\Big/ 5\sum_{i=1}^{10} m_i,$$

where m_i is the mean income of ith decile.

C. Kakwani's Method

This method was proposed in Kakwani and Podder [25]. This method uses a new coordinate system, which is shown in Figure A–11. The Lorenz curve is expressed by the new coordinate system as follows:

$$x = (F_i + Q_i)/\sqrt{2}, \qquad 0 < x < \sqrt{2}$$
$$y = (F_i - Q_i)/\sqrt{2}, \qquad 0 < y < 1/\sqrt{2}$$

where F_i and Q_i are cumulative household share and income share (see Figure A–11).

Table A-2
Estimates of Pareto Distribution

Year	Area	Parameter Estimate		Adjusted R-Squared
		$\ln A$	σ	
1969	Overall	18.800 (55.19)	−2.093 (−63.6)	0.9961
	Rural	21.590 (60.88)	−2.414 (−70.1)	0.9966
	Urban	19.252 (27.11)	−1.981 (−30.4)	0.9925
1975	Overall	17.360 (39.10)	−2.409 (−49.3)	0.9930
	Rural	18.341 (38.09)	−2.585 (−48.3)	0.9924
	Urban	18.582 (38.99)	−2.386 (−46.6)	0.9936
1981	Overall	17.744 (41.69)	−2.279 (−46.9)	0.9944
	Rural	18.821 (45.07)	−2.460 (−54.9)	0.9954
	Urban	20.147 (33.22)	−2.409 (−37.8)	0.9923
1986	Overall	14.690 (30.84)	−1.920 (−35.6)	0.9937
	Rural	16.082 (28.13)	−2.115 (−32.4)	0.9915
	Urban	15.886 (25.94)	−1.921 (−28.7)	0.9964

Source: Estimated by the author.
Note: Figures in the parentheses indicate t-value.

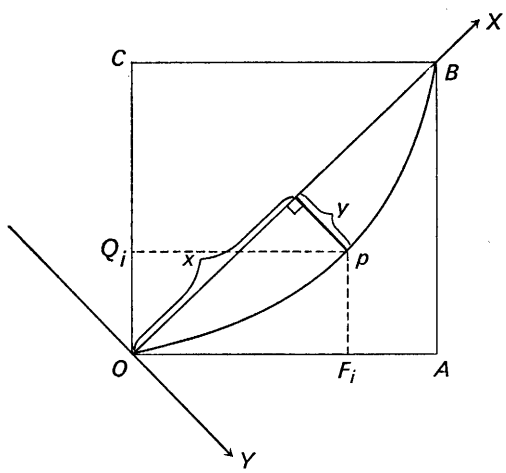

Figure A-11
New Coordinate System

Though various functional forms can be fitted to the Lorenz curve in the new coordinate system, Kakwani and Podder adopted the following form:

[5] $y = A \cdot x^{\alpha}(\sqrt{2} - x)^{\beta}$.

This form assumes that the Lorenz curve goes through the two points:

O and B in Figure A–11. The coefficients can be easily estimated by applying *OLS* to the logarithmic form of the above equation.

With these estimated coefficients the Gini coefficient can be calculated by integration and it is expressed as:

$$\text{Gini} = A\sqrt{2}^{(\alpha+\beta+3)}B(\alpha+1, \beta+1),$$

where $B(\alpha+1, \beta+1)$ is the beta function.

The Skewness of the Lorenz Curve

Kuznets concluded that "the greater inequality in developing countries was primarily a result of a high concentration of income in the top income group" and that "the share of the lower income groups was larger in the developing countries than in the developed countries," which means that "people in intermediate income groups in developing countries have a much smaller share of the total income than those in the same groups in developed countries" (Kakwani [24], p. 380). This was mentioned as regards the Lorenz curves of Figure 4–8 in chapter 4.

Figure A–12 shows these two types of Lorenz curve in the new coordinate system, I is for developing countries and II is for developed countries. Lorenz curve I is skewed toward O and Lorenz curve II is skewed toward B. The skewness can be measured by the value of x which brings about the highest value of y (expressed as x^* and x^{**} in Figure A–12). This value of x can be derived by differentiating the equation [5] with respect to x and setting the derivative equal to zero. And we obtain:

$$x^* = \frac{\sqrt{2}\alpha}{\alpha+\beta} = \sqrt{2} - \frac{\sqrt{2}}{(\alpha/\beta)+1}.$$

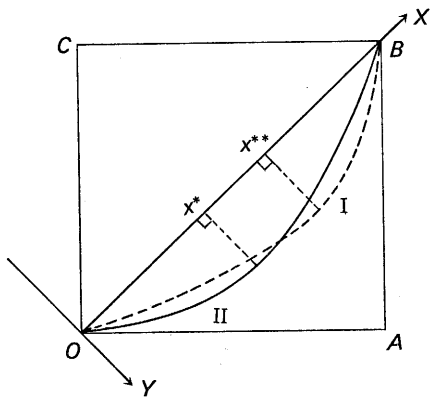

Figure A–12
Lorenz Curve for Developing and Developed Countries

This equation indicates that x^* is an increasing function of (α/β). Therefore, the smaller the ratio (α/β) is, x^* also is smaller and the skewness of the Lorenz curve moves toward that of the developed countries.

II. Decomposable Inequality Index

A. The Theil Index

The Theil entropy index is one of the decomposable inequality indices. The Theil index is expressed as:

$$[6] \quad T = \sum_i \sum_j \frac{y_{ij}}{Y} \ln \frac{y_{ij}/Y}{n_{ij}/N},$$

where n_{ij} is the absolute frequency of households of the ith group (for example, region) and the jth group (for example, income class), N is the total number of households $(=\sum_i \sum_j n_{ij})$, y_{ij} is the total income of the i–j group, Y is the total income of all households $(=\sum_i \sum_j y_{ij})$.

If we set the number of households of the ith group as N_i $(=\sum_j n_{ij})$ and the total income of the ith group as Y_i $(=\sum_j y_{ij})$ and the Theil index of the ith group as $T_i\left(=\sum_j \frac{y_{ij}}{Y_i} \ln \frac{y_{ij}/Y_i}{n_{ij}/N_i}\right)$, this equation can be decomposed as follows:

$$T = T_w + T_b,$$

where $T_w = \sum_i \left(\frac{Y_i}{Y}\right) T_i$ and $T_b = \sum_i \frac{Y_i}{Y} \ln \frac{Y_i/Y}{N_i/N}$.

T_w is the weighted average of the Theil index of group i, the weights being the income share of the group and called the "within-group component" or "within-component" in short. T_b is equal to the Theil index when there is no inequality within each group and each household receives the same level of income as the average income of the group. This is known as the "between-group component," or "between-component" in short. The ratio of T_w and T_b, to T, (T_w/T) and (T_b/T), is called the contribution of within-component and between-component, respectively.

B. Variance of Income Logarithm (Varlog)

The variance of income logarithm (varlog) is another decomposable inequality index. This index is expressed as:

$$V = \sum_i \sum_j \frac{n_{ij}}{N} (\ln m_{ij} - \bar{m})^2,$$

where m_{ij} is the mean income of the ith group and jth income class and m is the mean of income logarithm of all groups.

In the same manner as the Theil index, V can be decomposed as follows:

$$V = V_w + V_b,$$

where $V_w = \sum_i \left(\frac{N_i}{N}\right) V_i$ and $V_b = \sum_i \left(\frac{N_i}{N}\right) (\ln m_i - \bar{m})^2$.

V_w is a weighted average of the variance of income logarithm within group i, (V_i), the weight being the population share of the group (N_i/N) and it is called the "within-group component" or "within-component." T_b is the between-group variance of the income logarithm, which is equal to the variance of the income logarithm assuming income inequality within group i is absent, and is called the "between-group component" or "between-component." Their contribution to income inequality is defined as the ratio of T_w and T_b to the variance of income logarithm (V), that is, (V_w/V) and (V_b/V).

III. Decomposition of Gini Coefficient by Source of Income

The Gini coefficient can be decomposed by source of income (Rao [47]). In this section the methodology will be explained.

First of all we assume that the total income consists of n sources and therefore expressed as:

$$Y = \sum_{i=1}^{n} Y_i,$$

where suffix i indicates the ith source of income.

Now the Gini coefficient can be decomposed as follows:

$$\text{Gini} = \sum_{i=1}^{n} W_i \cdot G_i,$$

where W_i is the share of the ith source in the total income and G_i is the pseudo-Gini coefficient of the ith source.

The pseudo-Gini coefficient is calculated by applying the same method as is used for the Gini coefficient to the distribution of the ith source of income, which is ordered by the level of total income. Contrary to the Gini coefficient, the pseudo-Gini coefficient of the ith source of income uses the rank order of total income as a weight.

Therefore, if a source of income concentrates toward the lower income class in absolute terms, the pseudo-Gini coefficient will be negative (see

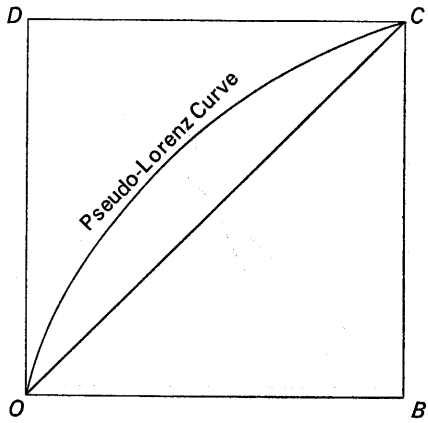

Figure A–13
Negative Pseudo-Gini Coefficient

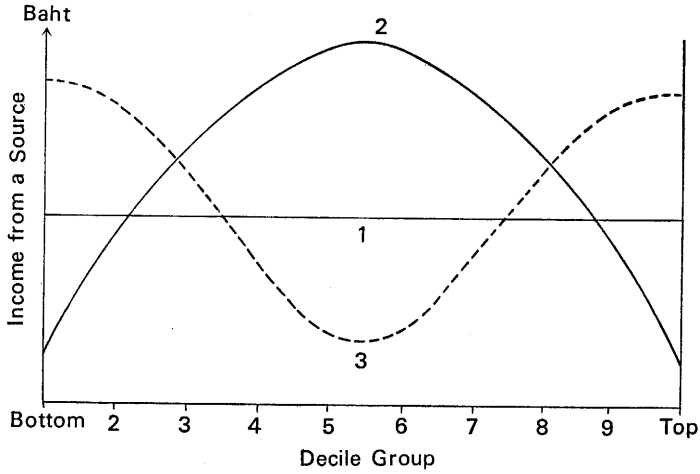

Figure A–14
Symmetric Distribution of Income of a Source

Figure A–13) and if it concentrates toward the higher income class the pseudo-Gini coefficient will be positive, which the Gini coefficient usually is.

It may be worthwhile to point out three cases of zero pseudo-Gini coefficient. One is the egalitarian case, that is, the case where every household has the same amount of income from the *i*th source, which is the same as the zero Gini coefficient (see line 1 in Figure A–14). The other cases are shown in Figures A–15 and 16. Figure A–15 is the case in which the in-

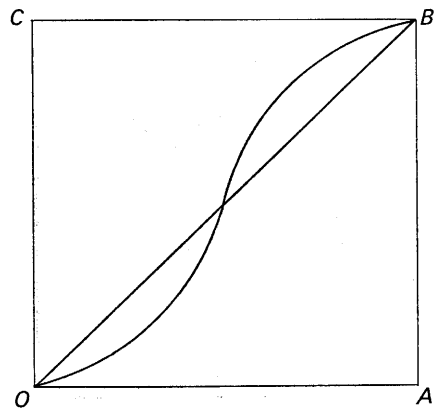

Figure A–15
Concentration toward the Middle-Income Class

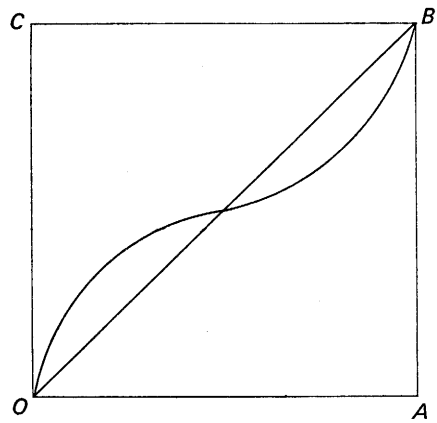

Figure A–16
Concentration toward Low- and High-Income Classes

come from the *i*th source concentrates toward the middle income class (see curve 2 in Figure A–14), and Figure A–16 is the case in which the income from the *i*th source concentrates toward both lower and higher income classes (curve 3 in Figure A–14). These three cases are cases of symmetry as shown in Figure A–14 and imply that the pseudo-Gini coefficient is not only an index of inequality but also an index of symmetry.